Active learning for computational chemogenomics

Research Article

For reprint orders, please contact reprints@future-science.com

Future Medicinal Chemistry

Active learning for computational chemogenomics

Aim: Computational chemogenomics models the compound–protein interaction space, typically for drug discovery, where existing methods predominantly either incorporate increasing numbers of bioactivity samples or focus on specific subfamilies of proteins and ligands. As an alternative to modeling entire large datasets at once, active learning adaptively incorporates a minimum of informative examples for modeling, yielding compact but high quality models. **Results/methodology:** We assessed active learning for protein/target family-wide chemogenomic modeling by replicate experiment. Results demonstrate that small yet highly predictive models can be extracted from only 10–25% of large bioactivity datasets, irrespective of molecule descriptors used. **Conclusion:** Chemogenomic active learning identifies small subsets of ligand–target interactions in a large screening database that lead to knowledge discovery and highly predictive models.

First draft submitted: 10 October 2016; Accepted for publication: 26 January 2017; Published online: 6 March 2017

Keywords: chemogenomics • computational chemistry and modeling • virtual screening

Daniel Reker[‡,1,2], Petra Schneider[1,3], Gisbert Schneider[1] & JB Brown*[,‡,4]

[1]Computer-Assisted Drug Design, Institute of Pharmaceutical Sciences, Department of Chemistry & Applied Biosciences, Swiss Federal Institute of Technology (ETH Zurich), Vladimir-Prelog-Weg 1-5/10, 8093 Zurich, Switzerland

[2]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, 500 Main St, Cambridge, MA 02139, USA

[3]inSili.com GmbH, Segantinisteig 3, 8049 Zurich, Switzerland

[4]Kyoto University Graduate School of Medicine, Center for Medical Education, Life Science Informatics Research Unit, Kyoto 606–8501, Japan

*Author for correspondence: jbbrown@kuhp.kyoto-u.ac.jp

[‡]Authors contributed equally

Background

Identifying new associations between small molecules and their macromolecular targets via computational prediction has been established in academic and industrial research workflows both for hit and lead discovery as well as for chemical biology [1–5]. The different approaches can be distinguished according to the origin of the data used for inferring new ligand–target relationships [6–8]. Receptor-based approaches extract information about the target on the level of the amino acid sequence, tertiary structure information or protein family relationships [9,10]. Conversely, ligand-based prediction methods rely on mathematical representation of ligand structures and comparisons guided by the chemical similarity principle (structurally similar ligands often exhibit similar bioactivity) [11–15]. The choice for either approach is strongly governed by data availability or simply personal preference, with no clear winner among the numerous retrospective comparisons or when reviewing the literature on prospective applications [16–18]. The benefit of using complementary approaches has been investigated previously and justifies the existence of a multitude of methods that have distinct applicability domains [19–21].

Computational chemogenomics (or proteochemometric modeling) is an integral part of the molecular informatics toolbox and represents a consequent coalescence of the ligand- and receptor-based philosophies [22–24]. Computational chemogenomic models leverage the information available by comparing the similarities of both ligands and targets simultaneously. Such developments were motivated by the completion of the human genome [25] and the successful application of consensus models [26–29], as well as the increasing value of pharmacological insight derived from investigating ligand–target networks [1,30].

FUTURE SCIENCE part of fsg

Chemogenomic data come to utility when, for instance, one has new compounds with a phenotypical readout but not target data, and wishes to generate hypotheses about the targets of the new compounds by comparing them to similar existing compounds for which target data are indeed known. Chemogenomic data come to utility in the target space as well when, for example, a team begins research into a new target and can begin by considering the existing bioactivity profiles of other similar proteins. Protein promiscuity [31] and the similarity principle [32] can provide directions in hit selection or lead optimization. The pair of perspectives for ligand-based and target-based hypothesis generation have been explored independently but typically not in combination. Intertwining information from both the ligand and the target space in a macroscopic, comprehensive model has been recognized as a possible strategy to benefit simultaneously from ligand and target similarities which thereby increases the domain of applicability compared to separate models [22,33].

Commonly, chemogenomics approaches rely on concatenating descriptors for ligands and targets as inputs to statistical learning methods that classify a query ligand–target pair as interacting or noninteracting [34,35]. Retrospective studies and a slowly rising number of prospective validations have shown that various machine learning techniques are indeed capable of navigating the interface of chemical and biological spaces [23,33]. They can efficiently use the available data to predict links between the provided sets of bioactive compounds and their biomacromolecular targets (termed 'Class I' prediction in [33], see Box 1 for terminology). Two noteworthy prospective applications of computational chemogenomics have applied the concept to discover ligands for simulated orphan targets [36] ('Class III' prediction) as well as study resistance development through point mutations in HIV non-nucleoside reverse transcriptase mutants [37]. These studies specifically highlight how computational chemogenomics methods are able to explore potential pharmacological relationships both from a ligand- as well as a target-centric perspective [38].

The ability to execute such methods is tightly coupled to the increasing data available through high-throughput and high-content screens. However, the cardinality of the theoretically available space generated through the product of all possible organic compounds and all biomacromolecular targets is computationally intractable, with estimates for the number of possible pairs reaching no less than 10^{65} [39]. This might serve in part as an explanation why most computational chemogenomic applications have focused on specific subfamilies of proteins and ligands [40]. Current complex machine learning models, for example following 'deep learning' principles [41,42], that are trained on comprehensive datasets are still difficult to interpret, and result in performances that are as of yet insignificant compared with existing methodologies. In particular, large ligand–target networks have been reported to suffer from a strong target selection bias [43], which may force models to learn historic target preferences rather than identifying informative ligand–target patterns [44].

While complex model research continues, active learning methods have recently gained attention in the drug discovery community [45,46]. In short, instead of fitting models to data *en masse*, active learning adds machine-picked examples in stepwise fashion, terminating once satisfactory prediction performance is achieved or once a specified number of data points have been included for model calculation (Figure 1 & Box 2). The machine learning model is re-trained after every data addition in order to adapt experimental design 'on-the-fly' for improved experimental efficiency. Prospective studies have recognized the ability of active learning to dynamically steer model development and rapidly identify structurally novel compounds for individual targets [46–51]. In such prospective chemistry applications, novel compounds would be predicted and assayed, after which those experimental results become the additional examples to learn from, and another round of modeling, prediction and validation begins. When applied retrospectively, active learning has also been recognized as a data mining technique to identify the most informative subset of data to fit high-quality models [52–55]. By starting from scratch, active learning can retrospectively pick only the examples necessary for model construction.

Driven by the importance of chemogenomics, and by the potential benefits of active learning, we set out to investigate whether chemogenomic active learning can identify key subsets of ligand–target pairs generated by large screening data of drug discovery projects in order to construct predictive models with reduced target bias and lend the subset of pairs selected to manual analysis.

Rarey and coworkers recently published an active learning framework for a multitarget problem in drug discovery, namely the ability of active learning combined with chemogenomic reasoning to build a predictive model for a target subfamily by using only a subset of available data [56]. Their investigation demonstrated efficient navigation of focused ligand–target spaces and provided a means for model training and experimental design.

The objective of our contemporaneous investigation challenges this hypothesis for much larger, full

Box 1. Chemogenomic prediction problems.

The goal of computational chemogenomics is to build predictive models by leveraging the similarity of compounds and similarity of proteins/targets. Brown *et al.* have previously defined four classes of prospective chemogenomic prediction problems [33], ordered in terms of increasing challenge. In Class I problems, the goal is to predict the missing values in a matrix of ligands and targets containing at least one data point per molecule. No new ligands or targets are used for prediction. In Class II, novel ligands are predicted for the existing targets. This class includes *de novo* design and lead optimization predictions. In Class III, new targets are predicted for existing ligands. This class includes screening for orphan, homolog or mutant proteins. In Class IV, neither the ligands nor the targets to be tested for association are included in the training/reference data, which is a stringent test of a model's ability to extrapolate from the reference bioactivity data

proteochemometric spaces. That is, we investigate the potential of active learning to act as the steering wheel for family-wide computational chemogenomic model building. We evaluate its ability to predict bioactivity and how it evolves a chemogenomic model [49,56–58], finding that protein/target family-wide chemogenomic active learning can build an interaction model from only a small fraction of bioactivity data points in a screening database. These models show high predictive performance on datasets many folds larger than that used for model construction. This leads to the implication that chemogenomic active learning might actually be able to computationally identify the most beneficial assays for subsequent execution and evaluation. It could serve as a platform to iteratively include the results in an actively updating model, which consequently would lead to making strides in improving discovery rates and reducing screening costs.

Materials

Compound–protein interaction data

We extracted ligand–target bioactivity from the ChEMBL SARfari databases [59], and from a recent G-protein coupled receptor (GPCR)-specific database GLASS [60] in which data were acquired from the database portal website in October 2015. The workflow to process the data is graphically demonstrated in Supplementary Figure 1.

SARfari databases were processed as follows. The databases contain bioactivity tables encoded as flat text files, with each interaction composed of a target domain (typically protein name), a compound ID, an assay type and a bioactivity record including the bioactivity metric (e.g., IC_{50}, EC_{50}, K_i) and relational value (e.g., = 10 μM or >1 nM). SARfari uses internal compound and protein IDs. We eliminated Starlite ADMET and Starlite functional assays, thereby retaining target-based biochemical and functional assays, with further restriction to exclusively human targets. Compound–protein interaction pairs (hereafter, 'CPIs') were further filtered to use K_i values for GPCRs (GPCR SARfari 3) and IC_{50} values for kinases (Kinase SARfari 5.01). 'Interactions' were extracted using a cutoff at 100 nM and 'noninteractions' were defined using a cutoff of 10 μM, thus separating the classes by two full logarithmic values. Interactions between the two ranges were discarded, as their classification is subjective. A postprocessing step was developed to eliminate any interactions with records in both the resulting interaction and noninteraction subsets. Additionally, interactions were eliminated for targets that did not contain activity data for at least 50 compounds.

The GPCR GLASS database is similarly available as a flat text file, using UniProt [61] IDs and InChI [62] keys as protein and compound identifiers, respectively. Using the UniProt IDs for targets, we reduced the GLASS database to human-specific proteins. GLASS was processed analogously to SARfari, however, due to the distribution of K_i bioactivities in GLASS, noninteractions were defined by the lower limit of 1 μM. A contradiction detection and 50 ligands/target filter was applied identical to that for SARfari datasets.

The resulting dataset sizes are given in Table 1. A number of analyses have shown that hit compounds turning out to be false positives contain common substructures, and in light of this, we have executed an analysis of the compounds in Table 1, flagging them by Rishton and Hann false-positive substructure flags [63,64], flagging them by a recent ligand multifamily promiscuity prediction tool [65], and finally, flagging them as potential pan-assay interference compounds (PAINS) [66] (see Box 3 for details on each flagging method). The prominent Hann false-positive flags were Michael acceptors, reactive alkylhalides, aliphatic methylene chains (n > 6) and disulfides. Note, however, that these compounds were not removed before modeling and evaluation, but were investigated for whether the learning strategies exhibit a bias toward selecting potential 'attrition' compounds. The raw distribution of ligands per target for each dataset is included as Supplementary Figure 2.

Molecule descriptors

Feature vectors to describe the interactions were constructed by concatenating the vector representations

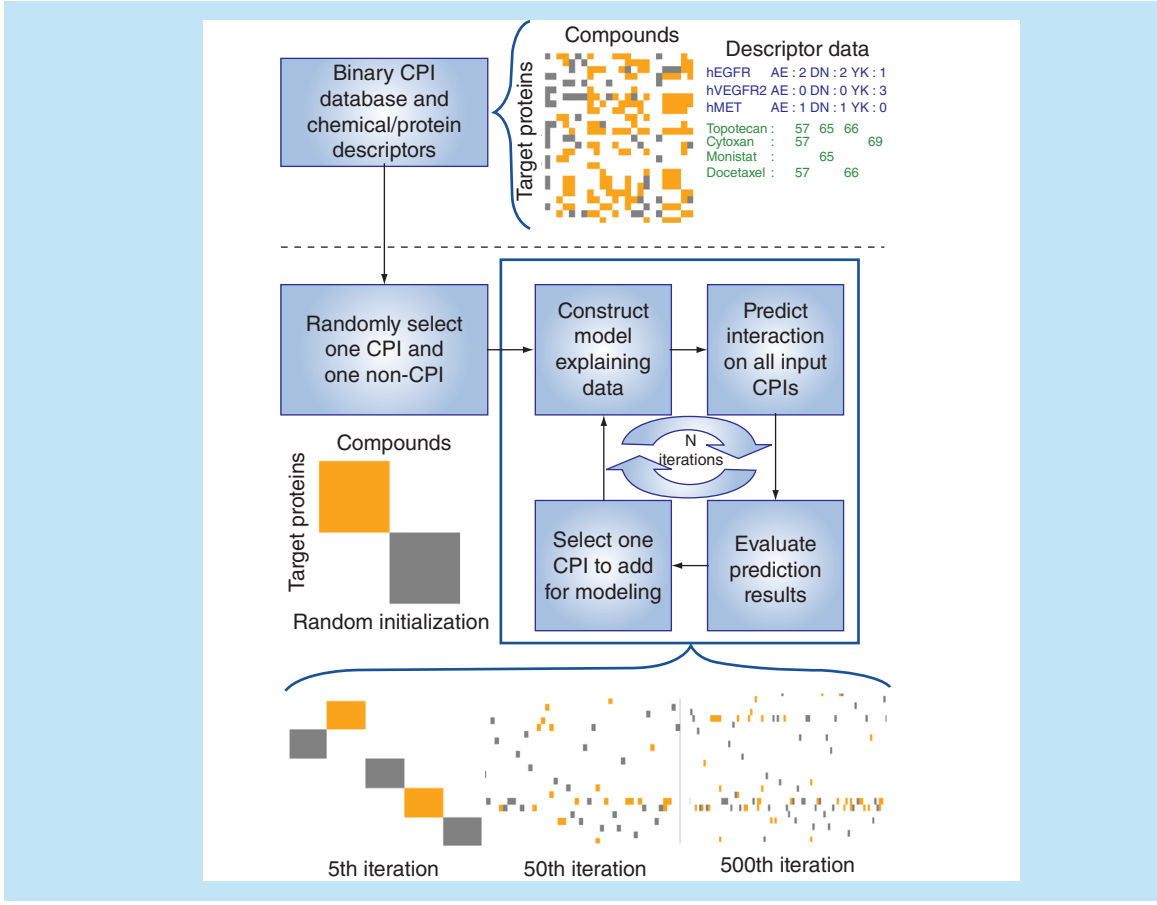


Figure 1. Concept of chemogenomic active learning. Chemogenomic active learning starts with an input dataset containing interactions and noninteractions, and molecule descriptors that represent features of compounds and proteins. After an initial random selection of one interaction and one noninteraction for generation of a minimal model, a model–predict–evaluate–incorporate cycle is executed. The cycle is repeated for a predetermined number of iterations, after which the modeling terminates. A goal of chemogenomic active learning is to terminate having extracted only the 10–20% of compound–protein pairs most informative for effective prediction on the remaining 80–90% of recorded bioactivity.
CPI: Compound–protein interaction

of each interaction's component compound and protein. To assess the sensitivity of results against molecular representation, multiple representations were evaluated.

For compounds, we computed the extended connectivity fingerprint (ECFP) [67] with a radius of 4 bonds hashed to 4096 bits (OpenEye OEChem library), as well as the 166-bit MACCS fingerprint [68]. For proteins, we applied the PROFEAT [69] protein descriptor to yield a 1497-dimensional vector describing the physicochemical and sequence properties of each target. In addition, we computed the frequency of dipeptides from the primary amino acid sequence of each protein (e.g., for the subsequence LDHLLLLAL in the prostaglandin D2 receptor with UniProt ID Q13258, the frequency of dipeptides would be LD = 1, DH = 1, HL = 1, LL = 3, LA = 1 and AL = 1).

Methods

Actively learned models

Datasets were randomly shuffled after input to avoid artifacts introduced through data ordering. Random forest models [70] (see Box 4) were trained using scikit-learn [71] by initializing RandomForestClassifier(n_estimators=500, max_features='sqrt') on one randomly picked interaction and one randomly picked noninteraction (Figure 1). From these sparsely trained models, active learning [45] was performed for 10,000 iterations, adding one new interaction to the model per iteration. In addition to the model methodology, the second and other crucial element of an active learning platform is the strategy used to select the next instance (CPI) to include in an updated model (Figure 1). Here, we implemented three CPI picking strategies (see Box 5 for a more detailed explanation of strategy semantics): random picking (simple random subsampling),

'greedy' picking (Equation 1, the CPI *i* which results in the maximum number of individual interaction classification trees *T(i)* that classify *i* as an interaction),

$$\text{argmax}_i \sum_{T \in Trees} \text{T(i)}, \text{T(i)} \in \{0,1\} \quad (1)$$

and 'curiosity' picking driven by predictive uncertainty (Equation 2, the CPI which results in the largest disagreement of individual tree predictions):

$$\text{argmax}_i \sum_{T \in Trees} [T(i) - F(i)]^2 \quad (2)$$

using

$$F(i) = [\sum_{T \in Trees} \text{T(i)}] / NumTrees$$

where F represents the average prediction over the whole forest of classification trees for CPI *i*, and where NumTrees corresponds to the number of trees in the forest.

Replicate experiments & evaluation

Input datasets were evaluated over 10 repeated executions of learning, where the execution number was used as the seed value for randomizing the input CPI list (and subsequent selection of the first two CPIs), and for seeding the random forest.

The actively learned CPI models were evaluated by multiple criteria, which are summarized in Table 2. First, we compared picking strategies by using the Matthews correlation coefficient (MCC) [72] of a model at each iteration of each execution for each strategy, where the full set of input CPIs was used for evaluation to ensure a comparability of the values obtained from different runs (see Box 6 for types of prediction outcomes and MCC concept). While this arguably biases the MCC results through adding known actives/inactives in the performance evaluation, this is a systematic bias equivalent for all evaluated runs and therefore does not influence our relative comparisons. The MCC is calculated as follows (Equation 3):

$$\text{MCC} = (\text{TP} * \text{TN} - \text{FP} * \text{FN}) / \sqrt{[(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})]} \quad (3)$$

The raw data from an MCC curve are jagged, and therefore, methods to smoothen the data and lend it to analytical methods can help in interpreting the results. We fitted MCC curve data to an exponential decay function $y(x) = a^* \exp^{(-b^*x)} + c$, where *a* controls the speed of growth, *b* controls the speed of decay and *c* is a constant used to position the decay curve. Substituting x for an iteration of learning and y for the MCC at that iteration, we fit (Equation 4):

$$MCC(Iter) = a^* \exp^{(-b^*Iter)} + c \quad (4)$$

by applying the SciPy [74] 'optimize' module and its member function 'curve_fit' (nonlinear least squares), after which we solved (Equation 5):

$$d\text{MCC}/d\text{Iter} = \text{v} \quad (5)$$

to find the iteration whose tangent line has slope v (see Results). The exponential decay function was chosen because its shape mirrors the shape expected by active learning – that is, sustained improvement in prediction performance up to a given limit, at which performance is saturated and little to no further improvement can be achieved through addition of more data. Solving the derivative (5) provides an estimate on the number of CPIs that can be added while maintaining a defined level of performance improvement.

In addition to the MCC values, we calculate the iterative true positive rate (TP/[TP+FN]) and the true negative rate (TN/[TN+FP]) for additional perspectives on evaluating model performances.

As shown in Table 1, chemogenomic datasets can be heavily biased toward actives. Therefore, picking behaviors were also evaluated by monitoring the ratio of picked interactions to noninteractions over the learning process. How chemical and protein spaces were dynamically explored was investigated by comparing each selected CPI instance to the previously selected CPI. This was done by measuring the similarity between the stepwise compounds (ECFP fingerprints with Tanimoto similarity) and stepwise proteins (using the Local Alignment Kernel [75]). Alternatively, the selected CPI instance was compared with the entire collection of previously selected CPIs to calculate the maximum similarity to the model's existing training set. Each type of comparison was analyzed using 2D histograms (heatmaps reflecting the binning of compound similarity on one axis and protein similarity on a second axis).

Next, 'target-trees', where each leaf of a tree represents an individual target and is augmented by a time-series heatmap reflecting that target's specific MCC trajectory, were constructed using neighbor join-

Box 2. Active learning for chemogenomics.

Despite advancements in combinatorial chemistry and high-throughput screening, it is prohibitively expensive to perform biochemical assays for target-level activity across all possible combinations of compounds and targets. In this respect, chemogenomic active learning seeks to achieve two goals. First, it seeks to yield the most informative compound–protein interactions in an existing bioactivity collection. Second, with sufficient predictive performance on the existing bioactivity collection, including an evaluation on those points not selected as the most informative, the method aims to predict the experimental validations most likely to succeed or provide valuable information before they are actually carried out, potentially saving a screening facility from incurring additional large costs

Table 1. Datasets in this study and their properties.

Data source	Kinase SARfari 5	GPCR SARfari 3	GPCR GLASS
Bioactivity type	IC_{50}	K_i	K_i
Threshold for interaction	100 nM	100 nM	100 nM
Threshold for noninteraction	10 μM	10 μM	1 μM
Resultant ligand–target pairs	39,706	47,602	69,960
Number of CPIs	19,231 (48%)	39,166 (82%)	49,815 (71%)
Number of non-CPIs	20,475	8436	20,145
Ligand statistics			
Number of ligands	20,897	31,751	44,484
Number of dual-class ligands	1292 (6%)	1427 (5%)	5302 (12%)
Number of Rishton flagged	2096 (10%)	1806 (6%)	2977 (7%)
Number of Hann flagged	696 (2%)	1270 (4%)	1648 (4%)
Number of promiscuity flagged	2695 (13%)	4041 (13%)	5986 (13%)
Number of PAINS flagged	5490 (26%)	8037 (25%)	11,269 (25%)
Target statistics			
Number of targets	98	100	110
Number of dual-class targets	98 (100%)	99 (99%)	82 (75%)

Datasets have been filtered from their original sources to select targets with at least 50 interaction data points after preprocessing. The GPCR datasets are heavily biased toward strongly binding ligands. Ligand–target pairs with bioactivity in between the defined ranges have been discarded.

ing as implemented in the BioPython package (version 1.65) [76], based on the protein–protein distance matrix generated by the local alignment kernel. Target trees were visualized using the ETE2 library (version 2.3.1) [77].

Reproducibility tests & comparison of means

For each picking strategy, we executed the Kolmogorov–Smirnov (KS) test for normality of results, where the samples to be tested are the average iterations (number of included CPIs) derived from the evaluation of (Equation 5). We used the implementation provided by the SciPy 'stats' module, called as stats.kstest(iterations,'norm').

Two-sided statistical t-tests, where mentioned below, were executed by calling stats.ttest_ind(group1, group2, equal_var=False) in SciPy.

Modeling of scrambled bioactivity end points

In certain situations, data may be easy enough to model such that nonsensical or randomized prediction end points can still be modeled with high accuracy. When such is successful, it raises doubt as to the validity of the original modeling, and when such fails, it signals that the original model indeed uncovered patterns in the data. This approach is known as *y*-scrambling, end point-shuffling or similar terminology. We performed an independent experiment for learning on the chemogenomic data that was preprocessed via *y*-scrambling in order to test the validity of the original unscrambled models.

External validation

To ensure that evaluation on predicting the entire dataset allows evaluating the learning behavior similar to a classical external validation test, we performed replicate external prediction using the GLASS dataset to build an actively learned model and the data points exclusively contained in the GPCR SARfari dataset as an external prediction set. The full SARfari dataset was predicted at every iteration of active learning on GLASS.

Results

Comparison of selection strategies

Ten executions each of CPI selection and modeling by using random, greedy and curiosity picking (Box 5) were evaluated with a fixed combination of ECFP and dipeptide descriptors. At each iteration, the MCC value of predictions on the entire input CPI set was computed to assess iterative model improvement, and the strategy's average performance, rather than any one individual run, was evaluated using the mean and standard deviation (Figure 2).

Inspection of performance across all datasets shows that the greedy selection strategy performs poorly in family-wise model development, with little to

Box 3. 'False-positive' chemical substructure flagging.

During the execution of quantitative high-througput screening, compounds might first be evaluated as hits, only to either be lost to attrition during optimization or to be later discovered to have interfered with the original assay readout. A manual analysis of these types of compounds has been researched, and several patterns have been suggested. From those suggestions, software tools have been implemented which check chemical structures for the potentially problematic substructures. The software tools 'flag' compounds when such substructures are detected. These tools can help guide the design of new ligand classes by recommending against the time and investment to screen such classes because they can be likely to fail downstream. In this article, four such flagging methods were used to assess the rate of potentially problematic structures in public data

no iterative improvement in terms of MCC performance. Random and curiosity pickers, on the other hand, show a marked iterative improvement in prediction accuracy. Curiosity selection exhibits the best performance of the three methods evaluated, and is the only method capable of achieving MCC values greater than 0.8 for all datasets at the limit of 10,000 selected CPIs. Hence, in our datasets comprising 40,000–80,000 CPIs, efficient selection of 25% or less of the CPIs can still yield a model with substantially high predictive MCC value over the remaining 75% or more of the data not used for model creation, which includes large numbers of compounds not used in model calculations.

The visual shapes of MCC curves show that a different learning speed, in terms of MCC improvement, is to be expected when performing active learning on different datasets and with different selection strategies (Figure 2). We aimed to quantify these qualitative differences and their statistical significance. Quantifying the ongoing speed of learning and the expected benefit from further learning can be done by computing the derivative (slope) of the fitted MCC learning curves (Equation 4), and solving it for a specified value representing learning speed (Equation 5). The solution, which is the corresponding iteration, then can be interpreted as the number of CPIs that can be included in a model while maintaining the specified rate of learning. In Table 3, results are listed when solving for (dMCC/dIter) slope values of 1.0 and 0.8. For both curiosity and random picking, only a few thousand CPIs were needed to converge on a model indicative of its predictive nature. However, compared with random selection, the curiosity picking method achieved clearly better average predictive performance at either stopping criteria.

By investigating for two slope values, we quantified the performance gain achieved compared with the additional number of required CPIs. In other words, we asked how much we would gain from potentially adding more complexity to the model by including more bioactivity data. Particularly in the case of the larger GLASS dataset, a 10% improvement in MCC ($0.56 \rightarrow 0.61$) was achieved by lowering the stopping criteria (to 0.8), resulting in addition of 500 extra (3000 total) CPIs for model development. From a larger perspective, this means that we achieved a 10% jump in performance by incorporating an extra 1% of the original data. We also note that adding CPIs to the random picker to reach $v = 0.8$ still resulted in lower performance than that of the curiosity picker at $v = 1.0$ in all datasets (Table 3), so even a much larger dataset selected at random might not be competitive with actively selected, smaller sets. In a follow-up, a Welch t-test between the curious and random picking methods at a fixed iteration of 2500 verified a sig-

Box 4. Semantics of decision trees and random forests.

- Decision trees are a data processing method which works by an analogy of human reasoning. Given a task, a decision tree examines particular features of data to see if there are threshold values of those features which, when iteratively considered, can separate the data into its different classes. In essence, a decision tree builds a set of 'if-then' rules (e.g., "if the number of hydrogen bond donors is greater than 3 then check X, else check Y."). At a 'leaf' layer in the decision tree, an if-then rule has the consequence of assigning a label (e.g. "this compound is toxic," to the object in question). After 'training' a decision tree, it can be used for examining new incoming cases of data and making predictions. This is analogous to a clinician who has gained experience in his specialty after examining many patients, and has mentally created a rough set of rules for diagnosing the causes of illnesses
- A random forest is then a collection of decision trees, where subsets of features to be considered for rule derivation are randomly selected for each tree in the forest. In this article, random forests are employed to evaluate subsets of the compound and protein descriptors, and to identify statistical patterns (decisions) that explain strong bioactivity or lack of bioactivity of compounds against different proteins

Box 5. Semantics of chemogenomic active learning compound–protein interaction picking strategies.

The key step to chemogenomic active learning is the compound–protein interaction (CPI) selection function that occurs at each iteration of modeling and update. We have employed random sampling as a baseline, which is uniform sampling from the input pool of CPIs. More interestingly, we have created two methods based on common strategies in computational decision making. In the 'greedy' strategy, the CPI that yields the highest score from the forest of decision trees is selected. Semantically, this means that 'greedy' stresses selection of interactions over noninteractions and picks a CPI which appears most likely to constitute a true interaction. In the 'curiosity' strategy, we select the CPI in which there is the least consensus among the decision trees when classifying the CPI. In other words, the random forest lacks understanding about the CPI because its constituent trees cannot come to a definitive conclusion about its interaction status, and therefore that the CPI warrants special attention during learning

CPI: Compound–protein interaction.

nificant difference between the MCC performance of the two methods ($p = 3.8 \times 10^{-20}$, 1.6×10^{-19} and 6.9×10^{-15} respectively for GLASS, GPCR SARfari and Kinase SARfari, ECFP/dipeptide descriptors).

We questioned reproducibility by asking if the per-method results across different executions are normally distributed. Using the mean iteration values in Table 3 for slope $v = 0.8$, we extracted the MCC values

Table 2. Chemogenomic active learning evaluation aspects.

Number	Aspect	Location in this article
1	Effect of CPI picking strategy	Figure 2 'Results', 'Additional Results', 'Discussion & Conclusion'
2	Ratio of actives to inactives selected	Figure 3 'Results', 'Discussion & Conclusion'
3	Compound and protein space explored	Figure 4, Supplementary Figure 3 'Results', 'Additional results', 'Discussion & Conclusion'
4	Individual target prediction performance	Figure 5, Supplementary Figure 4 'Results', 'Additional results'
5	Performance saturation (stopping criteria)	Table 3 'Results', 'Additional results', 'Discussion & Conclusion'
6	Result distribution for reproducibility	Table 4 'Results', 'Discussion & Conclusion'
7	Number of flagged compounds selected	Main text 'Additional results'
8	Effect of molecule descriptors on performance	Supplementary Figure 5 'Results', 'Additional results', 'Discussion & Conclusion'
9	Effect of metric used for evaluation	Supplementary Figure 6 'Additional results'
10	Effect of y-scrambling endpoints	Supplementary Figure 7 'Additional results'
11	Number of ligands per target during model evolution	Supplementary Figure 8 'Additional results'
12	Speed of covering target space	Supplementary Figure 9 'Additional results'
13	External prediction test	Supplementary Figure 10 'Discussion & Conclusion'

In order to comprehensively characterize chemogenomic active learning in terms of both metric-type performances as well as behavior-type patterns, multiple aspects were evaluated.
Locations in single quotation marks list where in the manuscript an aspect/result is discussed.
CPI: Compound–protein interaction.

at the dataset/picker-specific iteration and applied the KS test for normality. The results shown in Table 4 suggest that performances at stopping criteria are normally distributed, indicating that one could expect a specific range of results if further experiments were executed.

Selection ratio of interactions to noninteractions

We assessed the selection strategies for their tendencies in picking between interactions or noninteractions, as shown in Figure 3. Random selection converged on the original input distribution, an expected result. Greedy selection, while showing less favorable results until now in terms of predictive performance, dominantly selected active interactions regardless of the balance of the original dataset. For the curiosity selection method, a balanced selection of interactions and noninteractions was achieved, irrespective of whether the original dataset was balanced or not. This is notable for the highly imbalanced GLASS dataset containing approximately 70,000 CPIs in which 70% of them are interactions. Given the consistency of these findings in all datasets, we may attribute them as properties associated with respective pickers.

Compound & protein space exploration

To better understand the trajectories of the active learning strategies in chemical as well as biological space, we assessed similarities of the compounds and proteins for the CPIs selected during active learning (Figure 4).

First we considered the curiosity and random pickers for comparison of a selected compound compared with the previous iteration. Consecutive iterations predominantly selected dissimilar compounds in both methods. However, when we next considered the pickers for similarity between a selected compound versus all previously selected compounds, a difference in the methods emerged. We found that the random picker results in continuous selection of dissimilar chemical space, while the curiosity picker appeared to return to pockets of chemical space that were selected in prior iterations, as demonstrated by the higher compound similarities in the one-vs-all panels of Figure 4.

In terms of targets selected, curiosity and random methods demonstrated selection of the target family spaces (each containing ~100 targets) within 1000 iterations. However, the actual per-iteration movement in protein space shows that the curiosity selection can remain in a family subspace over multiple iterations more frequently than random selection, as evidenced by the deeper colored distribution at the high end of similarity values in Figure 4, and the expansion of Figure 4 to all executions, Supplementary Figure 3.

Predictive performance on the level of individual targets

We sought to further clarify how the iterative predictive improvement of the models through the picking strategies translates into improvement of predictive performance on the level of individual targets. Figure 5 compares 'target trees' for the three selection methods, where each leaf of a tree represents a single target, and the heatmap attached to each leaf reflects the time course of prediction performance (MCC) specifically with respect to that protein.

Per-target predictive performance on the Kinase SARfari dataset was good for the random and curiosity methods. For GPCRs, the results were less favorable for the random selection method. For the targets that are predictable, more iterations were needed compared to the curiosity picker. Particularly in the case of the larger GLASS dataset, the difference in target prediction performance is clear. Based on Table 3, we examined target trees by considering the per-target performance after only 3000 iterations of active learning. Here, the merit of curiosity selection is uncontested (inner curves of Figure 5).

Box 6. Prediction outcome types and Matthews' correlation coefficient.

- In two-class data modeling, we can arrive at four types of results: TPs, FPs, TNs and FNs. In chemogenomic modeling:
  - TPs represent interactions correctly predicted to be interactions
  - FPs represent noninteractions falsely predicted to be interactions
  - TNs represent noninteractions correctly predicted to be noninteractions, and
  - FNs represent interactions falsely predicted to be noninteractions
- FPs and FNs are respectively known as Type-I and Type-II errors in statistics. In a number of prediction assessment metrics such as the true positive rate, only one type of error is included in the calculation, which can yield deceptively high performance. In contrast to this, the Matthews correlation coefficient is a metric which includes all four types of prediction results into a single metric, and can handle biased or unbalanced data (see Equation 3 in the main text). The range of the Matthews correlation coefficient is between -1 and 1, where values greater than 0 indicate more correct predictions than combined Type-I and Type-II errors

FN: False negative; FP: False positive; TN: True negative; TP: True positive.

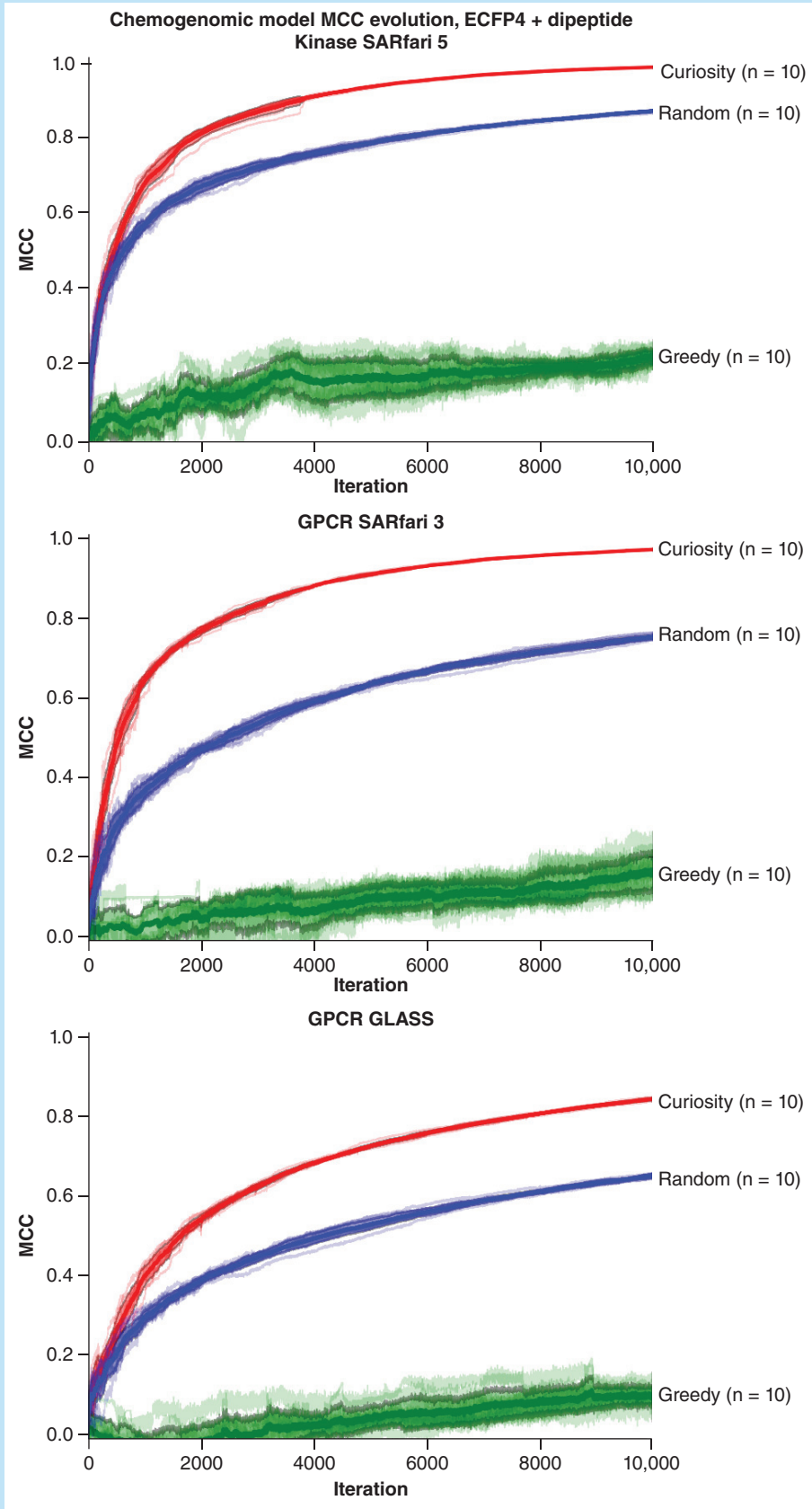

Chemogenomic model MCC evolution, ECFP4 + dipeptide
Kinase SARfari 5
MCC
Iteration
Curiosity (n = 10)
Random (n = 10)
Greedy (n = 10)
GPCR SARfari 3
GPCR GLASS
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10,000

Figure 2. Picking strategy comparison. The evolution of active learning MCC on the input datasets is evaluated for 10,000 iterations. n = 10 executions are used to statistically assess performance for different picking strategies. Raw results are lightly colored, the mean MCC per iteration is strongly colored using a bold line, and the standard deviation of performance is shown using transparent areas.
MCC: Matthews correlation coefficient.

Using greedy selection, a clear pattern emerges that a small cluster of targets is well predicted from early in the active learning, with no improvement on the remaining targets. Across multiple executions (Supplementary Figure 4), we find that certain target clusters are repeatedly predicted moderate to well, which suggests that the greedy selection is a local subfamily optimization method. This behavior was observed even when looking at target trees only up to 3000 iterations.

Impact of descriptor combination

To investigate whether the curiosity picker's performance can be achieved with different descriptions of ligands and proteins, we executed an evaluation of it on all combinations of ECFP/MACCS compound descriptors crossed with PROFEAT/dipeptide protein descriptors. As shown in Supplementary Figure 5, curiosity picking using any combination of one compound descriptor and one protein descriptor yielded similar results, with all selection strategies achieving similar MCC values at the 10,000 CPI cutoff. However, the MCC development within that span is different between the various description methods, with less pronounced initial MCC gains for MACCS+PROFEAT. This finding was consistent across all three investigated datasets. In all except this case, the initial learning behavior of any descriptor combination driven by curiosity picking clearly outperformed random selection as well as greedy selection (Figure 2 & Supplementary Figure 5). Curiosity-based experiments using tripeptide frequency were retrospectively executed, and predictive performance compared with dipeptide frequency was identical with respect to iterative MCC values (data not shown).

Table 3. Estimation of stopping iteration for compound–protein interaction instance selection methodologies.

Data: picking strategy	Solving *d*MCC/*d*Iter = 1.0		Solving *d*MCC/*d*Iter = 0.8	
	Mean ± std MCC 95% CI (2.5, 97.5)-%ile	Mean ± std iteration	Mean ± std MCC 95% CI (2.5, 97.5)-%ile	Mean ± std iteration
GLASS: random	0.33 ± 0.03 CI: 0.31–0.35 (0.28, 0.36)	1537 ± 272	0.39 ± 0.02 CI: 0.38–0.41 (0.36, 0.41)	2266 ± 211
GLASS: curiosity	0.56 ± 0.05 CI: 0.52–0.59 (0.45, 0.60)	2297 ± 479	0.61 ± 0.05 CI: 0.57–0.64 (0.50, 0.65)	2790 ± 549
GS3: random	0.45 ± 0.02 CI: 0.43–0.47 (0.41, 0.48)	1975 ± 240	0.51 ± 0.02 CI: 0.50–0.53 (0.48, 0.53)	2642 ± 177
GS3: curiosity	0.81 ± 0.01 CI: 0.81–0.81 (0.80, 0.82)	2294 ± 52	0.84 ± 0.01 CI: 0.83–0.84 (0.83, 0.84)	2591 ± 62
KS5: random	0.66 ± 0.02 CI: 0.64–0.67 (0.63, 0.68)	1974 ± 81	0.69 ± 0.014 CI: 0.68–0.70 (0.67, 0.71)	2377 ± 59
KS5: curiosity	0.83 ± 0.01 CI: 0.82–0.84 (0.81, 0.84)	2189 ± 78	0.86 ± 0.01 CI: 0.85–0.86 (0.84, 0.86)	2474 ± 102

Using the curves generated by calculating the MCC value at each iteration of active learning, we fitted curves to an exponential decay function and solved them for two values reflecting expectations about acceptable rates of model improvement. Solutions were analyzed for the stopping iteration and resulting MCC. Using the solved MCC values, the 95% CI was calculated by calling stats.t.interval in the SciPy package, and the 2.5- and 97.5-percentile values were calculated by the 'percentile' function in NumPy [73].
CI: confidence interval; MCC: Matthews correlation coefficient.

Table 4. Testing for Gaussian distribution of Matthews correlation coefficient results.

Data: picking strategy	Stopping iteration	Mean ± Std MCC	KS statistic	KS p-value
GLASS: random	2266	0.40 ± 0.01	0.14	0.99
GLASS: curiosity	2790	0.61 ± 0.01	0.14	0.99
GS3: random	2642	0.51 ± 0.02	0.20	0.84
GS3: curiosity	2591	0.81 ± 0.01	0.15	0.98
KS5: random	2377	0.69 ± 0.02	0.17	0.93
KS5: curiosity	2474	0.84 ± 0.01	0.27	0.40

Using the mean stopping iteration estimated by solving for dMCC/dIter = 0.8 (Table 3), raw MCC values were extracted and tested for normality by the KS test. High p-values signal that we cannot reject the null hypothesis that results are normally distributed.
KS: Kolmogorov–Smirnov; MCC: Matthews correlation coefficient.

Additional results

Different evaluation metrics & scrambling tests

We considered evaluation of results using other metrics such as the true positive rate, where positive points were ligands with strong inhibitory activity on a target. We found that use of True Positive Rate tended to over-estimate model performance [78] (see Supplementary Figure 6), particularly for the imbalanced GPCR datasets, and therefore emphasis on this metric has the possibility to mislead modelers and subsequent molecule designs. Instead, we observed that the true negative rate (TNR) was considerably lower in initial iterations of modeling, and that the TNR and MCC were strongly correlated over the course of active learning (Supplementary Figure 6). However, in many screening exercises, a majority of ligands do not bind to receptors, yielding large amounts of negative points. In these scenarios, overemphasis on the TNR would then also be misleading. In either scenario of imbalance, the MCC would provide a better estimate of overall prediction performance.

To test if our models were truly uncovering relationships in the ligand–target bioactivity databases, we executed an additional experiment in which the (non) interaction labels were scrambled before modeling. Experiments were again run in replicate to remove artifacts from any particular scramble. The results, shown in Supplementary Figure 7, show largely decreased MCC values for modeling on all three scrambled datasets. We conclude that results on original data cannot be attributed to chance and that the proposed methodology is successfully extracting informative CPIs from the input datasets to learn meaningful ligand–target relationships.

Target selection & space coverage

We questioned how many ligand associations to a target were included for modeling at a given iteration, and how these counts differed over the course of learning with respect to the picking strategy. As seen in Supplementary Figure 8, the number of ligands per target is widely distributed across many targets when using the random selection, as it simply samples interaction space and hence protein space in a uniform manner. This resembles the findings from observing Figure 4. Greedy selection, on the other hand, immediately fixes on single targets or small sets of targets, as the selection function (Equation 1) selects the CPI which yielded a maximum number of votes for interaction during prediction; hence the same target was repeatedly added to the training CPI set, and the performance on that single target grows as was reflected in Figure 5.

The evolutionary behavior for the curiosity picking method (Equation 2) is far more varied. First, we note the different behavior taken in different executions of experiments. For instance, in the GPCR SARfari dataset, two of the executions demonstrate an uncertainty about several targets, and construct models largely consisting of those targets and their ligands in initial iterations. However, using different random seeds to begin the learning with a different initial CPI, the other two executions demonstrate a balanced selection of ligands per target. This pair of patterns is also seen for the GPCR GLASS dataset.

After considering the MCC trees shown in Figure 5, we retrospectively analyzed how fast the target space was covered in each dataset, where coverage of a target means that the target has been selected at least once among all iterations of a particular execution. Although the rapid improvement of model performance in terms of MCC values for all targets might suggest that the curiosity picker would cover the target space the quickest, this was not the case (Supplementary Figure 9). Both curiosity and random selection always covered target space, but the iterations required to do so were different. In all executions, the iterations required by random selection were lower ($p = 0.003$, 0.024 and 0.001 for

the GLASS, GPCR SARfari and Kinase SARfari datasets respectively, ECFP/dipeptide descriptors, t-test). At the iterations (number of CPIs) derived in Table 3, the target spaces are largely covered by both strategies. For greedy selection, the target space was rarely fully covered, and when fully covered, the iteration at which all targets were included into the model was consistently at iteration 9000 or above. The greedy selection focuses strongly on individual targets to exploit the knowledge acquired about actives. This was visible in the target trees (Figure 5 & Supplementary Figure 4).

If we consider coverage of target space at iteration 2000 (Supplementary Figure 9), results are nearly identical for the curiosity and random pickers, yet there is a significant quantitative difference in prediction performance (Figure 2, Table 3 and t-test values). Therefore, since target space coverage is the same, it stands to reason that the chemical space being covered might be responsible for the difference in performance of the selection strategies.

Diversity & properties of selected compounds

The diversity of compounds included in a chemogenomic model will directly affect the amount of scaffold hopping achievable for prospective applications [36,79]. Figure 4 (2D histograms) and Supplementary Figure 3 demonstrate that the chemical space explored was dependent on the picking strategy. Curiosity selection drove the learning process to stay in the same region of compound space over more iterations than by random selection, yet the region could be left when a particularly challenging example was encountered in another region or when the region was sufficiently better understood than the remaining SAR data.

Given the difference in prediction performance after 1000 iterations of learning, particularly for the GPCR SARfari example (Figure 2), and given the difference in selection distribution at 1000 iterations (Figure 4), where target space coverage was mostly equivalent (see previous section), we attribute the performance difference in downstream iterations to the difference in compound selection. The difference in performance was smaller for the Kinase SARfari dataset (Figure 2), and interestingly, we see more of a resemblance in the histogram distributions (Supplementary Figure 3). Importantly, high MCC values, in particular for random picking, suggest that the Kinase SARfari set was relatively easier to model, which can potentially be explained by recurring binding motifs in the ligands that make binding and nonbinding easier to discriminate compared with GPCRs. However, we were not required to explicitly inject macro-level mechanistic perspectives in order to build predictive models.

As explained in the 'Methods' section and Box 3, various heuristics are available for suggesting when a hit compound could in fact be an 'attrition' compound. Using the Rishton, Hann, PAINS and promiscuity false positive flagging heuristics, a subsequent analysis of the first 2000 iterations of curiosity learning revealed that the compounds selected in these iterations had a flagging rate equal to that of the background input distribution (Table 1) regardless of picking strategy. This is reasonable given that Equations 1 & 2 are not equipped with any sort of chemical filters. Learning does not appear to be influenced implicitly through the potentially broad bioactivity profiles of these suggested 'attrition' compounds.

Discussion & conclusion

Model size & parameters

When Figure 2 and Table 3 are analyzed in tandem, replicate experiment has shown that curiosity selection of 1500 interactions and noninteractions each is sufficient for a moderately (MCC = 0.6) to highly (MCC = 0.8) predictive model of family-wide chemogenomic spaces. Curiosity selection was more efficient than random selection [54], fully corroborating the power of the active learning concept reported previously [46,49,56]. While random selection and curiosity-driven selection perform similarly in the first few hundred iterations, curiosity-based learning performance after 1000–1500 iterations is superior, and learning is maintained over a longer phase as seen when comparing slope values (Table 3). Through this, even much smaller, actively-selected datasets lead to better predictive models compared with models trained on larger, randomly sampled subsets. This performance includes all prediction classes in the early phase, and classes I and II in later phases (see Box 1), suggesting ligand discovery as at least one domain of applicability.

Compared with previous computational chemogenomic work where 100,000 or more interactions including presumed negatives were employed for model construction [33], the models generated herein are orders of magnitude smaller yet equally as efficient. A reduced need for experimentation has been identified as a major breakthrough achieved through active learning, with estimates of the required training data ranging from as low as 10% [56] to 30% [51] of the complete training set.

Importantly, while it has previously been pointed out that an optimal descriptor pair needs to be evaluated for a specific chemogenomics project [80] and cannot simply be inferred from individual descriptor performance on independent biological and chemi-

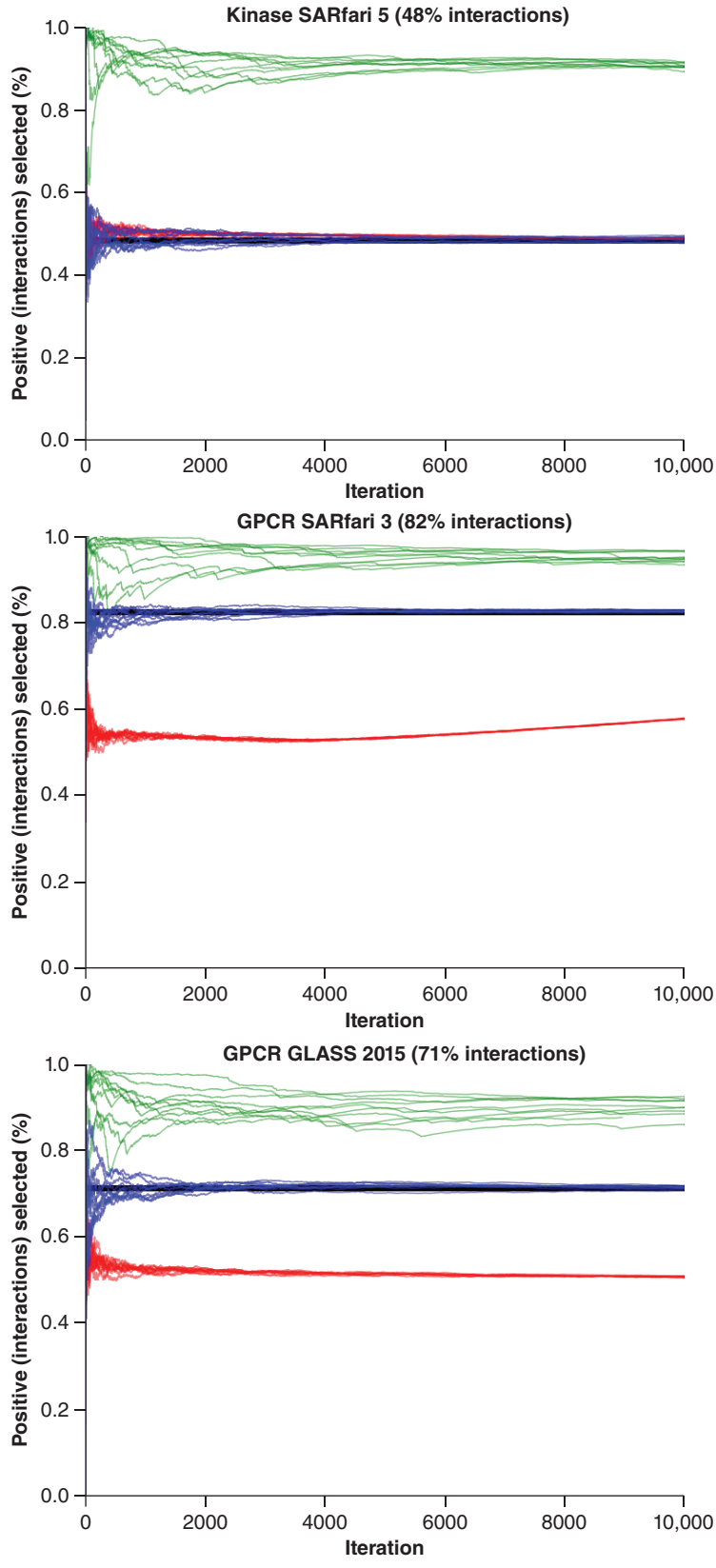

Kinase SARfari 5 (48% interactions)
Positive (interactions) selected (%)
1.0
0.8
0.6
0.4
0.2
0.0
0
2000
4000
6000
8000
10,000
Iteration
GPCR SARfari 3 (82% interactions)
Positive (interactions) selected (%)
Iteration
GPCR GLASS 2015 (71% interactions)
Positive (interactions) selected (%)
Iteration

Figure 3. Interaction to noninteraction ratio evolution. The ratio of interactions to noninteractions selected by each picking strategy over the course of active learning. A horizontal black line is placed at the original input dataset ratio, on which random selection imminently converged upon. Curiosity learning rapidly approaches selection of a balanced active-inactive training set, even in unbalanced datasets. Greedy selection, on the other hand, predominantly samples interactions. CPI selection ratio: Red: Curiosity; Blue: Random; Green: Greedy. CPI: Compound–protein interaction.

cal domains [33], the learning algorithms could select subsets of the underlying combination descriptors such that each subset yielded similar predictive performance. This would suggest that screening groups can derive a model that can be interpreted with the descriptor types they are comfortable with or those which have been developed for a specific purpose. While the speed of learning was influenced in terms of the underlying description, model performance at termination was similar using all possible combinations of descriptors. In other words, active learning can successfully navigate different projections of ligand-receptor spaces. Consistent with previous applications of active learning [58], we could show that family-dependent sequence similarities captured by the dipeptide descriptors can yield predictive chemogenomics models. However, applications that study the subtle pharmacology between similar proteins or multiple binding pockets might require alternative, more detailed descriptors that capture the structure of protein cavities relevant to ligand binding. We are undertaking a separate, comprehensive investigation of the properties of sequences that yield predictive models, and will report our analyses at a future date. On top of different description methods, implementations of the chemogenomic active learning concept using other modeling techniques are likely possible.

The most critical point that must be adhered to, however, is the provision of appropriate context for the model to drive the chemistry. Particularly in consideration of biological context, potential for compound optimization, as well as complementary 3D shape and electrostatics, domain-agnostic statistical algorithms will need contextually guiding hands to produce practical hit and lead molecules that will be tested by medicinal chemists. A joint effort by man and machine will likely be necessary to provide solutions to upcoming challenges in the pharmaceutical sciences [30,81].

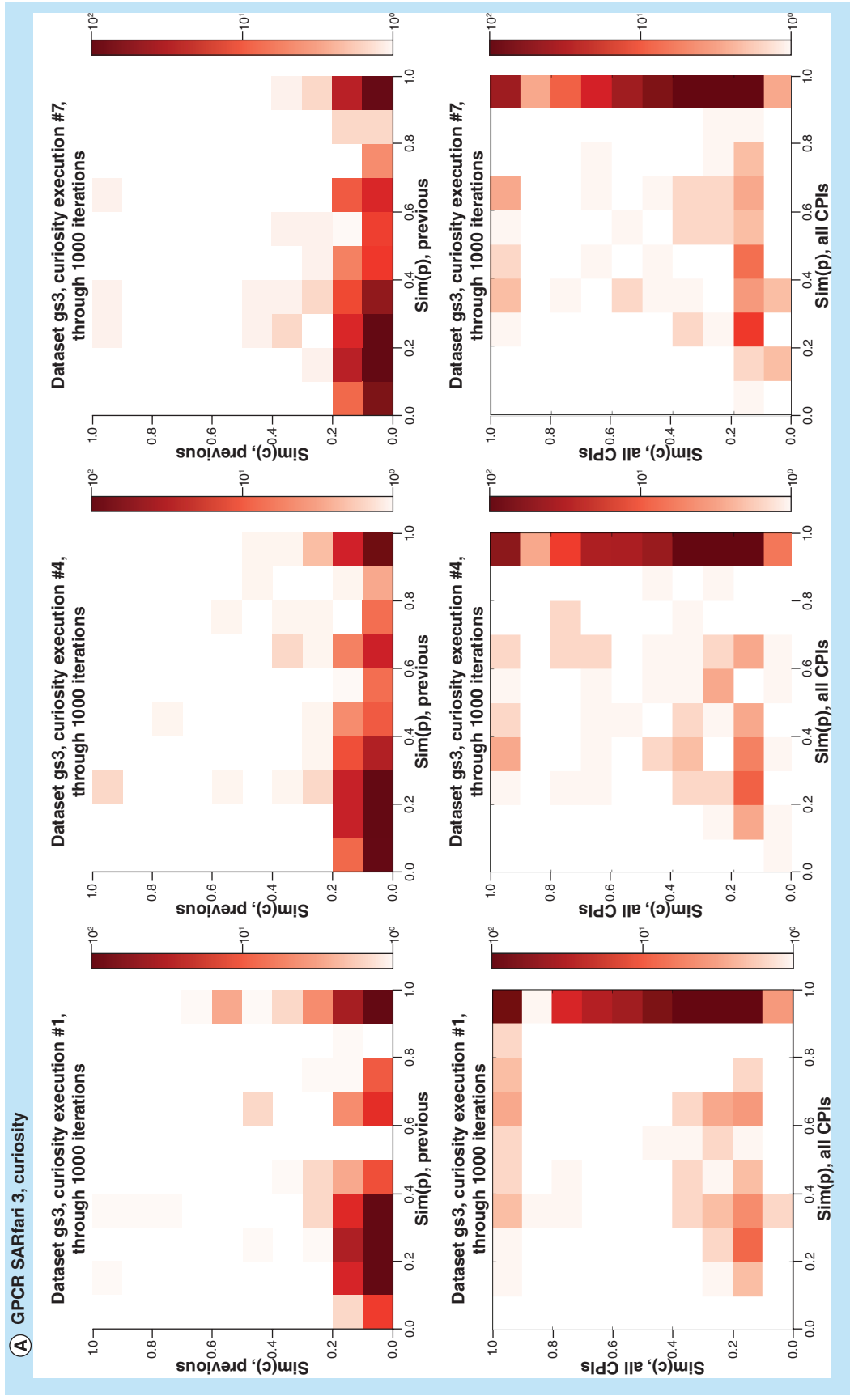
Balanced selection of interactions & noninteractions

We investigated the number of interactions and noninteractions picked by individual strategies. The greedy strategy, although not able to learn predictive family-wide models, sampled interactions 90% of the time. Apparently the hunt for actives can still be successful by relying on previously discovered actives without extrapolating the molecular patterns needed to understand the structure–activity relationship of the entire family investigated [82]. As expected, the random strategy sampled interactions and noninteractions according to the underlying class distribution in the dataset. The curiosity-based picking achieved a balanced selection of the two classes irrespective of the original data distribution. This result might serve at least in part as an explanation for the sustained MCC improvement of curiosity-learned models, compared with random picking.

Interestingly, a previous analysis on regression problems using CDK2 had suggested a more closely related sampling behavior of bioactivity between random sampling and curiosity-driven picking [45]. Further analysis will need to investigate whether such relation stems from the underlying bioactivity distribution in the utilized CDK2 data or whether this might be an inherent difference between active learning for classification versus regression problems.

Statistical tractability & external applicability

We have suggested to fit exponential decay functions on the learning performances of different selection strategies, which allowed us to identify certain stages of the learning process, and identify key iterations for different selection strategies. In fact thereby, the MCC curve slope combined with the number of iterations and the achieved model accuracy provides an important parameter to decide when to stop learning, and serves as input for automated switching strategies [83]. Other stopping and switching criteria have, for example, relied on external test sets [56], estimated model quality [83], or predictive uncertainty [84]. By statistically estimating achievable MCC values and required iterations, an informed decision on project development becomes possible for prospective applications – making active learning not only a competitive option but also a transparent method. The project team can use these interpretable metrics to decide on directions for model development and applications, for example to search for their desired intersection of CPIs and MCC by solving Equations 4 & 5. Taken together with the fact that we could show that statistical variations observed in learning behavior are normally distributed, we conclude that chemogenomic active learning is a statistically tractable and interpretable technique that can drive the analysis of existing screening library results as well as the design


A GPCR SARfari 3, curiosity
Dataset gs3, curiosity execution #1, through 1000 iterations
Dataset gs3, curiosity execution #4, through 1000 iterations
Dataset gs3, curiosity execution #7, through 1000 iterations
Sim(c), previous
Sim(c), all CPIs
Sim(p), previous
Sim(p), all CPIs
0.0
0.2
0.4
0.6
0.8
1.0
10^2
10^1
10^0

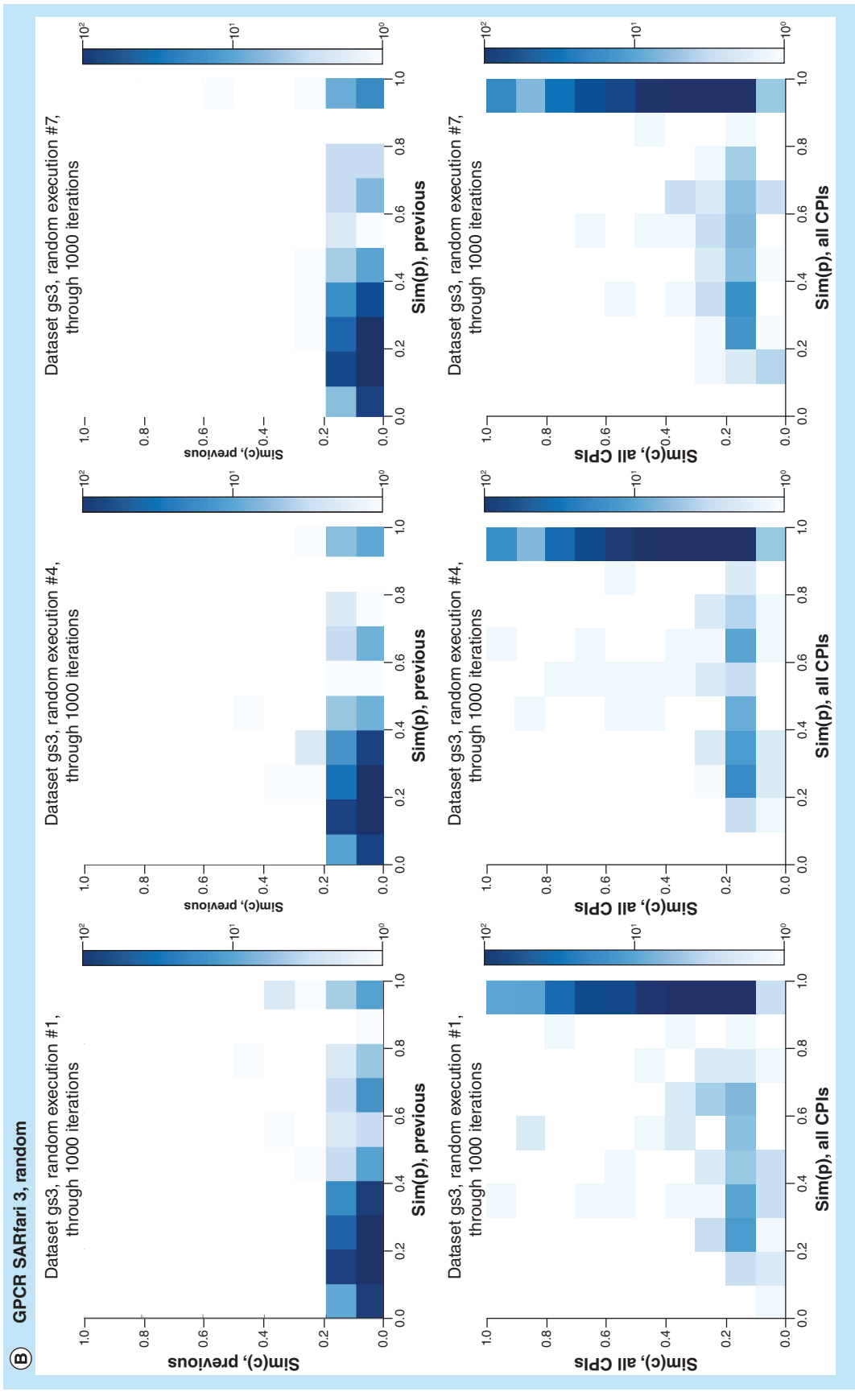


Figure 4. Compound-protein space exploration distributions. The comparison of a newly added CPI relative to its previously selected CPI is shown on two axes (first and third row). Horizontal axes represent protein similarity and vertical axes represent compound similarity. The comparison of a CPI relative to the entire existing training CPI set is similarly given (second and fourth row). Left to right, executions 1, 4 and 7 on GPCR SARfari are shown for the curiosity **(A)** and random (B) pickers for the first 1000 iterations of active learning. CPI: Compound–protein interaction.

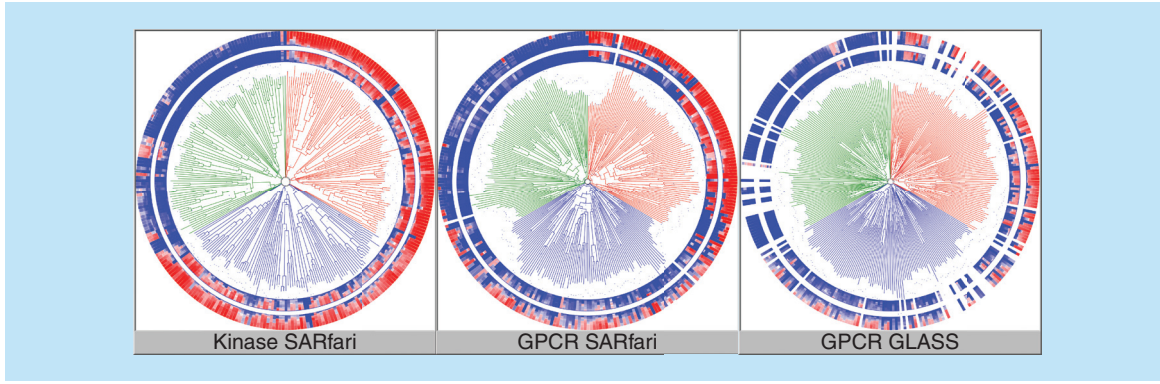


Figure 5. Per-target Matthews correlation coefficient trees. The per-target Matthews correlation coefficient (MCC) values (heatmaps; blue: low MCC; red: high MCC) calculated over the course of active learning are affixed to the family tree. For rendering purposes, mean MCC values are binned at every 100 iterations of learning. If a target contains only interactions, no MCC calculation is possible, and therefore, no heatmap is visualized. Tree colors correspond to the same color scheme of Figure 1 (Red: Curiosity; Blue: Random; Green: Greedy). The outer heatmaps indicate the MCC performance evolution over all 10,000 iterations, while the inner circle shows the development at early learning stages during the first 3000 iterations.

of experiments through computational pattern recognition, yet the framework provides sufficiently flexibility to be monitored and adaptively fine-tuned by a project team.

An external prediction of the GPCR SARfari dataset yielded surprisingly satisfactory MCC values that were strikingly similar to those achieved during the active learning of GLASS (Supplementary Figure 10). This suggests an applicability of learned models to completely novel compounds that have not been considered during learning. An additional investigation is needed to fully define the domain of applicability for purely external prospective applications.

Implications & future directions

The methods and results herein provide a tantalizing prospect that chemogenomic models of greatly reduced size can be efficiently created for any given CPI database. Critically, this will require additional work on larger databases, such as STITCH [85], the SARfari superset ChEMBL [59] or PubChem [86]. Orthogonally, many protein families have yet to be explored, such as the nuclear hormone receptor family or proteases. In a different stratification, we might ask how well the curiosity method performs for datasets exclusively representing different stages of chemical biology or drug discovery, namely the levels of screening hits, clinical trials and marketed drugs.

Prospective validation studies of reduced-complexity models built by chemogenomic active learning are critical. As a number of recent studies have indeed validated that the computational chemogenomic concept can lead to prospective discovery of interactions [36–37,87–88], we anticipate that actively learned models will be capable of similar novel discovery [48,49,51]. Given the increasing applicability of chemogenomics to uncover untested ligand–target pairs, many different exciting applications come to mind. For example, environmental agencies may consider applying computational chemogenomics as a way to generate hypotheses about the effect of pollutants generated during manufacturing processes [89]. In another application, deorphanization of natural products used in cancer therapy [90,91] can provide the starting CPIs to initiate an actively learned chemogenomic model which generates testable hypotheses of new drug–target interactions in uncharacterized drugs for specific cancer cell lines, such that the results of tested hypotheses are fed back into the model for subsequent hypothesis generation.

It was previously shown that active learning could query separate ligand- or target-based models to aid in improving the understanding of polypharmacological networks [58]. We have extended this hypothesis using chemogenomic modeling to capture the combined ligand–target space and aim at extrapolating knowledge from the interaction patterns. Our results suggest that equal numbers of ligands per target are not required for building chemogenomic models when examples are picked such that they benefit the understanding of family interaction space as a whole. Curiosity selection has shown that in certain cases it remains focused on either a specific target or a small group of similar targets. It would then appear that in certain cases, curiosity selection is building local SAR models for specific targets in spurts. The idea of many per-target quantitative structure–activity relationship models as a chemogenomic model has been explored previously [21,92–96]. A key difference between these

per-target approaches and the approach explored here is that we have removed the requirement to have a sufficient number of ligands per target in the per-target models, under the presumption that a sufficient number of similar ligand–target pairs also have similar bioactivity.

Adaptively trained models are expected to be a competitive option for driving pharmaceutical hit discovery to identify compounds with desired target profiles, and for risk control during development through early discovery of off-target interactions [47,97–98]. Put another way, the implication of our findings is that when models built by chemogenomic active learning on existing data are coupled with experimental screening platforms such that cycles of predicting, experimenting and model updating are iteratively performed, the potential reduction in experimental labor, time, and cost is large [99,100].

Future perspective

Having shown herein that a small subset of a ligand–target database is sufficient for predicting bioactivity on the entire collection, companies and screening centers with collections of tens of thousands of compounds screened against panels of targets can apply the active learning concept in order to extract knowledge about the key CPIs necessary for structure–activity relationship (SAR) understanding, and to prospectively screen additional libraries against the resulting ensembles of reduced-complexity models. Groups equipped with the necessary infrastructure will iteratively execute cycles of model–predict–experiment–incorporate. Beyond the human GPCR and kinase results here, groups can test the concept on other key families and organisms.

Acknowledgements

An academic license from OpenEye Scientific Software was used for chemical data processing, descriptor calculations and PAINS evaluation. Software developed by the open-source software community was indispensable in executing this research. Related discussions with Y Okuno of Kyoto University and J Bajorath of the University of Bonn were helpful. The authors also wish to thank the reviewers of the article for constructive criticism.

Financial & competing interests disclosure

G Schneider and P Schneider are the founders of inSili.com LLC, Zurich. This work was financially supported by ETH Zurich and Kyoto University. Exclusive compute time on a ProLiant DL580 Generation 9 server donated from Hewlett-Packard Enterprise Japan is kindly acknowledged. Support from the Japan Society for the Promotion of Science was used for computational resources under a Grant-in-Aid for Young Scientists (B) 25870336 and a Grant-in-Aid for Scientific Research (S) JP16H06306 (to JB Brown). D Reker is grateful for support from the Swiss National Science Foundation (P2EZP3_168827 to D Reker). This research was supported by a grant from the OPO-Foundation, Zurich (to G Schneider). This work was also supported by the JSPS Core-to-Core Program (A: Advanced Research Networks, to JB Brown, G Schneider). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Ethical conduct of research

The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

Supplementary data

To view the supplementary data that accompany this paper, please visit the journal website at: www.future-science.com/doi/full/10.4155/fmc-2016-0197

Executive summary

- Modeling of large ligand–receptor bioactivity databases often done using entire database, but benefit of extra data not proven.
- Intelligent selection of key ligand–receptor pairs can reduce the size of data needed to 5–25% of original data.
- Reduced selection still yields high predictability on target families and individual targets.
- Chemistry selected by intelligent selection is more useful for modeling than chemistry selected by random sampling.
- Proposed compound–protein selection and model method is statistically tractable and reproducible.

References

1 Hopkins AL. Network pharmacology. *Nat. Biotechnol.* 25(10), 1110–1111 (2007).

2 Stahl M, Guba W, Kansy M. Integrating molecular design resources within modern drug discovery research: the Roche experience. *Drug Discov. Today.* 11(7–8), 326–333 (2006).

3 Schneider P, Schneider G. *De novo* design at the edge of chaos. *J. Med. Chem.* 59(9), 4077–4086 (2016).

4 Brown JB, Okuno Y. Systems biology and systems chemistry: new directions for drug discovery. *Chem. Biol.* 19(1), 23–28 (2012).

5 Lee AY, St Onge RP, Proctor MJ *et al.* Mapping the cellular response to small molecules using chemogenomic fitness signatures. *Science* 344(6180), 208–211 (2014).

6 Lavecchia A, Cerchia C. *In silico* methods to address polypharmacology: current status, applications and future perspectives. *Drug Discov. Today* 21(2), 288–298 (2016).

7 Koutsoukas A, Simms B, Kirchmair J *et al.* From *in silico* target prediction to multi-target drug design: current databases, methods and applications. *J. Proteomics.* 74(12), 2554–2574 (2011).

8 Jacoby E. Computational chemogenomics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 1(1), 57–67 (2011).

9 Lyne PD. Structure-based virtual screening: an overview. *Drug Discov. Today* 7(20), 1047–1055 (2002).

10 Berger SI, Iyengar R. Network analyses in systems pharmacology. *Bioinformatics* 25(19), 2466–2472 (2009).

11 Geppert H, Vogt M, Bajorath J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* 50(2), 205–216 (2010).

12 Achenbach J, Tiikkainen P, Franke L, Proschak E. Computational tools for polypharmacology and repurposing. *Future Med. Chem.* 3(8), 961–968 (2011).

13 Ekins S, Mestres J, Testa B. *In silico* pharmacology for drug discovery: methods for virtual ligand screening and profiling. *Br. J. Pharmacol.* 152(1), 9–20 (2007).

14 Vidal D, Garcia-Serna R, Mestres J. Ligand-based approaches to in silico pharmacology. In: *Chemoinformatics and Computational Chemical Biology.* Bajorath J (Ed.). Humana Press, Totowa, NJ, 489–502 (2011).

15 Brown JB, Urata T, Tamura T, Ara MA, Kawabata T, Akutsu T. Compound analysis via graph kernels incorporating chirality. *J. Bioinform. Comput. Biol.* 8(Suppl. 01), 63–81 (2010).

16 Ripphausen P, Nisius B, Peltason L, Bajorath J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J. Med. Chem.* 53(24), 8461–8467 (2010).

17 Tan L, Geppert H, Sisay MT, Gütschow M, Bajorath J. Integrating structure- and ligand-based virtual screening: comparison of individual, parallel, and fused molecular docking and similarity search calculations on multiple targets. *ChemMedChem* 3(10), 1566–1571 (2008).

18 Sams-Dodd F. Target-based drug discovery: is something wrong? *Drug Discov. Today* 10(2), 139–147 (2005).

19 Krüger DM, Evers A. Comparison of structure- and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors. *ChemMedChem* 5(1), 148–158 (2010).

20 Wilson GL, Lill Ma. Integrating structure-based and ligand-based approaches for computational drug design. *Future Med. Chem.* 3(6), 735–750 (2011).

21 Reker D, Rodrigues T, Schneider P, Schneider G. Identifying the macromolecular targets of *de novo*-designed chemical entities through self-organizing map consensus. *Proc. Natl Acad. Sci. USA* 111(11), 4067–4072 (2014).

22 Bajorath J. A perspective on computational chemogenomics. *Mol. Inform.* 32(11–12), 1025–1028 (2013).

23 van Westen GJP, Wegner JK, Ijzerman AP, van Vlijmen HWT, Bender A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *MedChemComm* 2(1), 16–30 (2011).

24 Bleicher KH. Chemogenomics: bridging a drug discovery gap. *Curr. Med. Chem.* 9(23), 2077–2084 (2002).

25 Caron PR, Mullican MD, Mashal RD, Wilson KP, Su MS, Murcko MA. Chemogenomic approaches to drug discovery. *Curr. Opin. Chem. Biol.* 5(4), 464–470 (2001).

26 Drwal MN, Griffith R. Combination of ligand- and structure-based methods in virtual screening. *Drug Discov. Today Technol.* 10(3), e395–e401 (2013).

27 Tanrikulu Y, Krüger B, Proschak E. The holistic integration of virtual screening in drug discovery. *Drug Discov. Today* 18(7), 358–364 (2013).

28 Givehchi A, Schneider G. Multi-space classification for predicting GPCR-ligands. *Mol. Divers.* 9(4), 371–383 (2005).

29 Koch CP, Perna AM, Weissmüller S *et al.* Exhaustive proteome mining for functional MHC-I ligands. *ACS Chem. Biol.* 8(9), 1876–1881 (2013).

30 Butcher EC, Berg EL, Kunkel EJ. Systems biology in drug discovery. *Nat. Biotech.* 22(10), 1253–1259 (2004).

31 Nobeli I, Favia AD, Thornton JM. Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.* 27(2), 157–167 (2009).

32 Maggiora GM, Johnson MA. *Concepts and Applications of Molecular Similarity*. Wiley, NJ, USA (1990).

33 Brown JB, Niijima S, Okuno Y. Compound–protein interaction prediction within chemogenomics: theoretical concepts, practical usage, and future directions. *Mol. Inform.* 32(11–12), 906–921 (2013).

34 Brown JB, Okuno Y, Marcou G, Varnek A, Horvath D. Computational chemogenomics: is it more than inductive transfer? *J. Comput. Aided Mol. Des.* 28(6), 597–618 (2014).

35 Lapinsh M, Prusis P, Lundstedt T, Wikberg JES. Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. *Mol. Pharmacol.* 61(6), 1465–1475 (2002).

36 Yabuuchi H, Niijima S, Takematsu H *et al.* Analysis of multiple compound–protein interactions reveals novel bioactive molecules. *Mol. Syst. Biol.* 7, 472 (2011).

37 van Westen GJP, Wegner JK, Geluykens P *et al.* Which compound to select in lead optimization? Prospectively

validated proteochemometric models guide preclinical development. *PLoS ONE* 6(11), e27518 (2011).

38 Cortes-Ciriano I, Van Westen GJP, Lenselink EB, Murrell DS, Bender A, Malliavin T. Proteochemometric modeling in a Bayesian framework. *J. Cheminform.* 6, 35 (2014).

39 Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* 16(1), 3–50 (1996).

40 Cortes-Ciriano I, Murrell DS, Van Westen GJ, Bender A, Malliavin TE. Prediction of the potency of mammalian cyclooxygenase inhibitors with ensemble proteochemometric modeling. *J. Cheminform.* 7, 1 (2015).

41 Pérez-Sianes J, Pérez-Sánchez H, Díaz F. Virtual screening: a challenge for deep learning. In: *10th International Conference on Practical Applications of Computational Biology & Bioinformatics.* Saberi Mohamad M, Rocha PM, Fdez-Riverola F, Domínguez Mayo JF, De Paz FJ (Eds). Springer International Publishing, Cham, Switzerland, 13–22 (2016).

42 Gawehn E, Hiss JA, Schneider G. Deep learning in drug discovery. *Mol. Inform.* 35(1), 3–14 (2016).

43 Hakes L, Pinney JW, Robertson DL, Lovell SC. Protein–protein interaction networks and biology – what's the connection? *Nat. Biotechnol.* 26(1), 69–72 (2008).

44 Mestres J, Gregori-Puigjané E, Valverde S, Solé RV. Data completeness – the Achilles heel of drug-target networks. *Nat. Biotechnol.* 26(9), 983–984 (2008).

45 Reker D, Schneider G. Active-learning strategies in computer-assisted drug discovery. *Drug Discov. Today.* 20(4), 458–465 (2015).

46 Murphy RF. An active role for machine learning in drug development. *Nat. Chem. Biol.* 7(6), 327–330 (2011).

47 Besnard J, Ruda GF, Setola V *et al.* Automated design of ligands to polypharmacological profiles. *Nature* 492(7428), 215–20 (2012).

48 Desai B, Dixon K, Farrant E *et al.* Rapid discovery of a novel series of Abl kinase inhibitors by application of an integrated microfluidic synthesis and screening platform. *J. Med. Chem.* 56(7), 3033–3047 (2013).

49 Reker D, Schneider P, Schneider G. Multi-objective active machine learning rapidly improves structure–activity models and reveals new protein–protein interaction inhibitors. *Chem. Sci.* 7, 3919–3927 (2016).

50 Naik AW, Kangas JD, Langmead CJ, Murphy RF. Efficient modeling and active learning discovery of biological responses. *PLoS ONE* 8, e83996 (2013).

51 Naik AW, Kangas JD, Sullivan DP, Murphy RF. Active machine learning-driven experimentation to determine compound effects on protein patterns. *Elife* 5, e10047 (2016).

52 Ueno T, Rhone TD, Hou Z, Mizoguchi T, Tsuda K. COMBO: an efficient bayesian optimization library for materials science. *Mater. Discov.* 4, 18–21 (2016).

53 Wei K, Bilmes J, Edu RUW, Edu BUW. Submodularity in data subset selection and active learning. *Proc. 32nd Int. Conf. Mach. Learn.* 37, 1954–1963 (2015).

54 Alvarsson J, Lampa S, Schaal W *et al.* Large-scale ligand-based predictive modelling using support vector machines. *J. Cheminformatics 2016.* 8(1), 948–962 (2016).

55 Clark JH, Frederking R, Levin L. Toward active learning in data selection: automatic discovery of language features during elicitation. Presented at: *Sixth International Conference on Language Resources and Evaluation.* Marrakech, Morocco, 28–30 May 2008.

56 Lang T, Flachsenberg F, Von Luxburg U, Rarey M. Feasibility of active machine learning for multiclass compound classification. *J. Chem. Inf. Model.* 56(1), 12–20 (2016).

57 Ahmadi M, Vogt M, Iyer P, Bajorath J, Fröhlich H. Predicting potent compounds via model-based global optimization. *J. Chem. Inf. Model.* 53(3), 553–559 (2013).

58 Kangas JD, Naik AW, Murphy RF *et al.* Efficient discovery of responses of proteins to compounds using active learning. *BMC Bioinformatics* 15(1), 143 (2014).

59 Bento AP, Gaulton A, Hersey A *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42, D1083–D1090 (2014).

60 Chan WKB, Zhang H, Yang J *et al.* GLASS: a comprehensive database for experimentally-validated GPCR-ligand associations. *Bioinformatics* 31, btv302 (2015).

61 UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* 43, D204–D212 (2015).

62 Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC international chemical identifier. *J. Cheminform.* 7, 23 (2015).

63 Rishton GM. Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discov. Today* 8(2), 86–96 (2003).

64 Hann M, Hudson B, Lewell X, Lifely R, Miller L, Ramsden N. Strategic pooling of compounds for high-throughput screening. *J. Chem. Inf. Comput. Sci.* 39(5), 897–902 (1999).

65 Schneider P, Rothlisberger M, Reker D, Schneider G. Spotting and designing promiscuous ligands for drug discovery. *Chem. Commun. (Camb.)* 52, 1135–1138 (2015).

66 Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* 53(7), 2719–2740 (2010).

67 Rogers D, Hahn M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50(5), 742–754 (2010).

68 Symyx. MACCS Structural Keys. San Ramon, CA: MDL Information Systems Inc.; 2005.

69 Li Z-R, Lin HH, Han LY, Jiang L, Chen X, Chen YZ. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.* 34(Suppl. 2), W32–W37 (2006).

70 Breiman L. Random forests. *Mach. Learn.* 45(1), 5–32 (2001).

71 Pedregosa F, Varoquaux G, Gramfort A *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830 (2011).

72 Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405(2), 442–451 (1975).

73 Van Der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation (2011). https://arxiv.org/abs/1102.1523

74 Jones E, Oliphant T, Peterson P *et al.* SciPy (2001). www.scipy.org/

75 Saigo H, Vert J-P, Ueda N, Akutsu T. Protein homology detection using string alignment kernels. *Bioinformatics* 20(11), 1682–1689 (2004).

76 Cock PJA, Antao T, Chang JT *et al.* BioPython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11), 1422–1423 (2009).

77 Huerta-Cepas J, Dopazo J, Gabaldon T. ETE: a Python environment for tree exploration. *BMC Bioinformatics* 11(1), 24 (2010).

78 Todeschini R, Ballabio D, Grisoni F. Beware of unreliable Q^2! A comparative study of regression metrics for predictivity assessment of QSAR models. *J. Chem. Inf. Model.* 56(10), 1905–1913 (2016).

79 Schneider G, Neidhart W, Giller T, Schmid G. "Scaffold-Hopping" by topological pharmacophore search: a contribution to virtual screening. *Angew. Chemie Int. Ed. Engl.* 38(19), 2894–2896 (1999).

80 Brown JB, Niijima S, Shiraishi A, Nakatsui M, Okuno Y. Chemogenomic approach to comprehensive predictions of ligand–target interactions: a comparative study. In: *2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops*. IEEE, PA, USA, 136–142 (2012).

81 Schneider G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discov.* 9(4), 273–276 (2010).

82 Bajorath J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* 1(11), 882–894 (2002).

83 Donmez P, Carbonell JG, Bennett PN. Dual strategy active learning. In: *Machine Learning: ECML 2007.* Springer, 116–127 (2007).

84 Bloodgood M, Vijay-Shanker K. A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009).* Association for Computational Linguistics, CO, USA, 39–47 (2009).

85 Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* 44, gkv1277 (2015).

86 Kim S, Thiessen PA, Bolton EE *et al.* PubChem substance and compound databases. *Nucleic Acids Res.* 44(D1), D1202–D1213 (2016).

87 Wang F, Liu D, Wang H *et al.* Computational screening for active compounds targeting protein sequences: methodology and experimental validation. *J. Chem. Inf. Model.* 51(11), 2821–2828 (2011).

88 Cheng F, Liu C, Jiang J *et al.* Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.* 8(5), ep-value2503 (2012).

89 Nicolotti O, Benfenati E, Carotti A *et al.* REACH and in silico methods: an attractive opportunity for medicinal chemists. *Drug Discov. Today* 19(11), 1757–1768 (2014).

90 Schneider G, Reker D, Chen T, Hauenstein K, Schneider P, Altmann K-H. Deorphaning the macromolecular targets of the natural anticancer compound doliculide. *Angew. Chemie Int. Ed. Engl.* 55(40), 12408–12411 (2016).

91 Reker D, Perna AM, Rodrigues T *et al.* Revealing the macromolecular targets of complex natural products. *Nat. Chem.* 6(12), 1072–1078 (2014).

92 Wassermann AM, Dimova D, Bajorath J. Comprehensive analysis of single- and multi-target activity cliffs formed by currently available bioactive compounds. *Chem. Biol. Drug Des.* 78(2), 224–228 (2011).

93 Lounkine E, Kutchukian P, Petrone P, Davies JW, Glick M. Chemotography for multi-target SAR analysis in the context of biological pathways. *Bioorganic Med. Chem.* 20(18), 5416–5427 (2012).

94 Yao Z-J, Dong J, Che Y-J *et al.* TargetNet: a web service for predicting potential drug-target interaction profiling via multi-target SAR models. *J. Comput. Aided. Mol. Des.* 30(5), 413–424 (2016).

95 Keiser MJ, Setola V, Irwin JJ *et al.* Predicting new molecular targets for known drugs. *Nature* 462(7270), 175–181 (2009).

96 Lipinski C, Hopkins A. Navigating chemical space for biology and medicine. *Nature* 432(7019), 855–861 (2004).

97 Hopkins AL, Mason JS, Overington JP. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* 16(1), 127–136 (2006).

98 Schneider G, Reker D, Rodrigues T, Schneider P. Coping with polypharmacology by computational medicinal chemistry. *Chim. Int. J. Chem.* 68(9), 648–653 (2014).

99 Reutlinger M, Rodrigues T, Schneider P, Schneider G. Combining on-chip synthesis of a focused combinatorial library with computational target prediction reveals imidazopyridine GPCR ligands. *Angew. Chemie Int. Ed. Engl.* 53(2), 582–585 (2014).

100 Rodrigues T, Schneider P, Schneider G. Accessing new chemical entities through microfluidic systems. *Angew. Chemie Int. Ed. Engl.* 53(23), 5750–5758 (2014).